Misclassification in Case-Control Studies of Gene-Environment Interactions: Assessment of Bias and Sample Size

Montserrat Garcia-Closas,[1]
Nathaniel Rothman, and Jay Lubin

Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda Maryland 20892

Abstract

In studies of gene-environment interactions, exposure misclassification can lead to bias in the estimation of an interaction effect and increased sample size. The magnitude of the bias and the consequent increase in sample size for fixed misclassification probabilities are highly dependent on the prevalence of the misclassified factor and on the interaction model. This paper describes a relatively simple approach to assess the impact of misclassification on bias in the estimation of multiplicative or additive interactions and on sample size requirements. Applications of this method illustrate that even small errors in the assessment of environmental or genetic factors can result in biased interaction parameters and substantially increased sample size requirements that can compromise the feasibility of the study. Also, an example is provided where nondifferential misclassification biases an additive interaction parameter away from the null value, even under conditions where a multiplicative interaction parameter will always be biased toward the null value. Efforts to improve the accuracy in measuring both genetic and environmental factors are critical for the valid assessment of gene-environment interactions in case-control studies.

Introduction

Measurement error in exposure assessment is one of the major sources of bias in epidemiological studies. Most discussions on the effects of misclassification of exposure have focused on the impact on the relative risk and sample size in studies of a single factor (1–9). In contrast, less attention has been given to the influence of misclassification on the assessment of interactions between two or more factors (4, 10). In a recent paper, García-Closas *et al.* (11) showed that under a set of conditions often satisfied in studies of gene-environment interactions, both differential and nondifferential misclassification of a binary environmental factor biases a multiplicative interaction effect toward the null value. This result is also true for misclassification of genetic factors. As a result of misclassification, the required sample size to detect a departure

from the null hypothesis of no multiplicative interaction with a given statistical power will be increased. The impact of misclassification in the study of additive interactions is more difficult to predict and less well understood.

The increase in sample size for fixed misclassification probabilities is dependent on the prevalence of the environmental and genetic factors and on the type and magnitude of the interaction being evaluated (4, 10). Because studies to detect interactions typically require large sample sizes (12–16), further increases in sample size due to exposure misclassification could compromise the feasibility of the study (10). The evaluation of the effects of misclassification at the study design phase allows investigators an opportunity to consider alternative measures of exposure with different levels of accuracy and to identify situations where high-quality exposure assessment is crucial. The objective of this paper is to describe a relatively simple approach to quantify the impact of misclassification on bias in the estimation of interaction effects and on the required sample sizes. In the next sections, we describe and illustrate the approach with examples.

Materials and Methods

Consider a case-control study designed to investigate the presence of an interaction between a genetic and an environment factor. Environmental factor is broadly defined as endogenous or exogenous risk factors such as weight, endogenous levels of hormones, and cigarette smoking. For simplicity, assume that both the environmental ($E = e$) and genetic ($G = g$) factors are binary variables that take values of 1 for exposed or susceptible and 0 for unexposed or nonsusceptible. Disease status ($D = d$) takes the values of 1 for affected and 0 for unaffected. The odds ratio OR_{eg} measures the association between disease and the environmental and genetic factors. Relative to subjects not exposed to the environmental or genetic factor, we define the following odds ratios (Table 1): OR_{10} denotes the odds ratio for nonsusceptible subjects exposed to the environmental factor; OR_{01} denotes the odds ratio for susceptible subjects not exposed to the environmental factor; and OR_{11} denotes the odds ratio for susceptible subjects exposed to the environmental factor.

The multiplicative interaction parameter, Ψ, is defined as the ratio of the joint odds ratio and the product of the odds ratios for each factor at the reference level of the other factor, namely, $\Psi = OR_{11}/(OR_{10} \times OR_{01})$. In the absence of a multiplicative interaction, $\Psi = 1.0$ and $OR_{11} = OR_{10} \times OR_{01}$.

The additive interaction parameter, Φ, is defined as the ratio of the joint excess risk ($OR_{11} - 1$) and the sum of the excess risks for each factor at the reference level of the other factor, namely, $\Phi = (OR_{11} - 1)/(OR_{10} - 1) + (OR_{01} - 1)$. Other definitions for additive interaction parameters are pos-

Table 1 Definition of odds ratios (OR_{10}, OR_{01}, OR_{11}) and interaction parameters (Ψ and Φ)[a] for the relationship between two dichotomous environmental and genetic factors and disease

		Genetic factor (G)	
		$G = 0$	$G = 1$
Environmental factor (E)	$E = 0$	1.0^b	OR_{01}
	$E = 1$	OR_{10}	OR_{11}

^{a} $\Psi = \dfrac{OR_{11}}{OR_{10} \cdot OR_{01}}$

$\Phi = \dfrac{(OR_{11} - 1)}{(OR_{10} - 1) + (OR_{01} - 1)}$

^{b} Reference category.

sible (17) but will not be discussed in this paper. In the absence of an additive interaction, $\Phi = 1.0$ and ($OR_{11} - 1$) = ($OR_{10} - 1$) + ($OR_{01} - 1$). It should be noted that Φ is undefined when both OR_{10} and OR_{01} are 1.0, and that whereas Ψ takes values from 0 to $+\infty$, Φ takes values from $-\infty$ to $+\infty$.

Misclassification of a dichotomous exposure is defined by the misclassification probabilities sensitivity (*se*) and specificity (*sp*; Ref. 18). Sensitivity is the probability that a truly exposed subject is classified as exposed, and specificity is the probability that a truly unexposed subject is classified as unexposed. Nondifferential misclassification occurs when the misclassification probabilities are independent of the disease status, whereas differential misclassification occurs when the misclassification probabilities are dependent on the disease status. In the examples presented in this paper, we assume nondifferential misclassification; however, the approach described in this section can be used for both nondifferential and differential misclassification. Because nearly all instruments in epidemiology have some degree of error, sensitivity and specificity can also be defined for two instruments with different degrees of accuracy rather than for an error-free and an error-prone instrument. We will refer to the more accurate instrument as "gold standard" and the less accurate instrument as "error prone".

Sample Size Calculations. Sample size calculations presented in this paper were performed using the approach described by Lubin and Gail (19) and discussed by García-Closas and Lubin (16). These calculations can be performed using the program POWER that is available free of charge by e-mail from connorj@mail.nih.gov. In the examples presented in the next section, calculations assumed independence of the environmental and genetic factors in the population, a two-sided type I error of 5%, a type II error of 20% (*i.e.,* power = 80%), a case:control ratio of 1:1, and a rare disease in the population (defined in the examples as $P(D = 1|E = 0, G = 0) = 0.001$).

To calculate the sample size required to detect a multiplicative interaction of magnitude Ψ or an additive interaction of magnitude Φ, values for OR_{10} and OR_{01} need to be specified. These parameters are often difficult to specify, and the marginal odds ratios for the environmental factor (OR_E) and genetic factor (OR_G), *i.e.,* the odds ratios for each factor when the other factor is ignored, are often better known. The relationship between OR_{10}, OR_{01}, and Ψ (or Φ) and the marginal odds ratios (OR_E and OR_G) is given in "Appendix 1". Sample size calculations in two of the examples of multiplicative interactions presented in this paper are based on estimates of marginal

effects from previous studies and Ψ, rather than on OR_{10}, OR_{01}, and Ψ.

An Approach to Assess the Impact of Misclassification on Bias and Sample Size. The impact of misclassification of binary and independent factors, measured with errors that are independent of each other, can be assessed by the following procedure:

(*a*) Specify values for $P(E = 1)$, $P(G = 1)$, OR_{10}, OR_{01}, and OR_{11} in the absence of misclassification or when using a "gold standard" instrument.

(*b*) Calculate the required sample size for a given power to detect the interaction effect Ψ or Φ.

(*c*) Calculate $P(E^* = 1)$, $P(G^* = 1)$, OR^*_{10}, OR^*_{01}, and OR^*_{11} for values of sensitivity and specificity of the environmental and genetic factors as indicated in "Appendix 2", where "*" denotes the observed parameters in the presence of misclassification or when using an "error prone" instrument.

(*d*) Calculate the sample size using the observed parameters in the presence of misclassification.

It should be noted that this methodology is not applicable to ordered categorical or continuous exposure variables because without restrictions on the disease rate and on the form of the odds ratio function, the shape of the relationship with disease will generally be distorted by the measurement error (20).

Effect of Misclassification on Bias and Sample Size. Both differential and nondifferential misclassification of environmental or genetic factors bias a multiplicative interaction effect toward the null value, provided that the environmental and genetic factors are binary and independent, errors are independent, and the sum of sensitivity and specificity is ≥ 1 (*i.e.,* the classification instrument is better than random) (11). Under these circumstances, the sample size required to reject the null hypothesis of no multiplicative interaction with a given statistical power will be increased.

The direction of the bias to the additive interaction parameter in the presence of misclassification is more difficult to predict because we do not have a general rule as in the case of multiplicative interactions (11). Using the method described in the previous section, we explored empirically the direction of the bias to the additive interaction parameter, Φ, under a range of parameter values, assuming the same conditions indicated above for a multiplicative interaction. We found that under these conditions, nondifferential misclassification of the genetic or environmental factor generally tends to bias the additive interaction parameter toward the null value. However, we did find several examples where the additive interaction is biased away from the null in the presence of nondifferential misclassification in the environmental or genetic factor assessment. Although most of these scenarios were extreme situations, we found examples that can be encountered in practice. These examples followed a pattern where a protective factor measured with reduced specificity interacts with a risk factor of disease. We illustrate this situation in an example presented in Table 5. However, the approach described in this section can be used to assess the direction of the bias in each particular situation.

Results

Example of a 2-fold Multiplicative Gene-Environment Interaction. Consider an example of a multiplicative gene-environment interaction where the odds ratio for the effects of the environmental factor alone (OR_{10}) and the genetic

Table 2 Minimum number of cases (case : control ratio = 1) required to detect a 2-fold multiplicative gene-environment interaction (OR_{10} = 2.0, OR_{01} = 2.0, OR_{11} = 8.0) for different levels of accuracy of the environmental and genetic factors[a]

| Environmental factor sensitivity | Genetic factor sensitivity | Prevalences | | $\Psi*$[b] | Odds ratios | | | No. of cases |
		Environmental factor	Genetic factor		OR_{10}	OR_{01}	OR_{11}	
1.0	1.0	0.50	0.50	2.00	2.00	2.00	8.00	720
0.8	1.0	0.40	0.50	1.56	1.71	2.57	6.86	1600
1.0	0.95	0.50	0.48	1.83	2.18	1.91	7.64	900
0.8	0.95	0.40	0.48	1.46	1.82	2.39	6.38	2044

[a] Specificity for both genetic and environmental factor assessment = 1.0.
[b] $\Psi*$, observed interaction parameter.

factor alone (OR_{01}) are both 2.0, and the joint odds ratio for both factors (OR_{11}) is 8.0. Because the joint odds ratio is two times what would be expected under a multiplicative risk model, these values represent a 2-fold interaction (Ψ = 2.0). This example corresponds to a pattern of interaction where both the genetic and the environmental factors increase the risk of disease by themselves, and the joint effect is different from the effect of each factor acting alone [pattern 4 as described by Khoury *et al.* (21) and model E as described by Ottman *et al.* (22)]. We chose an example of this pattern because we believe that it is reasonable in the context of complex multifactorial diseases like cancer, where environmental and genetic factors are likely to influence the risk of cancer through multiple pathways.

Table 2 illustrates the impact of reducing sensitivity of the environmental factor assessment from 1.0 to 0.80, both in the absence and presence of reduced sensitivity in the assessment of the genetic factor (from 1.0 to 0.95). Although measures of genetic markers are generally considered less prone to error than measures of environmental exposures, some degree of error may be present due to technical errors in determining the genotype or due to failure to analyze or identify relevant alleles (8, 10). In Table 2, the prevalence for both factors is 0.5, and the specificity for the assessment of both the genetic and environmental factors is 1.0.

In the absence of misclassification of the genetic factor, reducing sensitivity of the environmental factor assessment from 1.0 to 0.80 increases the sample size from 720 to 1600 (2.2-fold; Table 2). This increase in sample size is driven by changes in the observed prevalence of the environmental factor and the observed odds ratios. In this example, the interaction parameter is underestimated from 2.00 to 1.56, the effect of the environmental factor alone is underestimated from 2.0 to 1.71, and the joint effect is underestimated from 8.00 to 6.86. In contrast, the effect of the genetic factor alone is overestimated from 2.00 to 2.57, although we assumed no errors in the genetic factor assessment. This bias occurs because the genotype effect on disease is larger among truly exposed than truly unexposed subjects; therefore, when truly exposed subjects are wrongly classified as unexposed because of a reduced sensitivity of exposure, the observed genotype effect among subjects classified as unexposed will be biased away from the null. The same amount of exposure error when sensitivity of the genetic factor is 0.95 rather than 1.00 further increases the sample size from 1600 to 2044. Thus, errors in exposure assessment coupled with errors in measuring the genetic susceptibility factor can have a substantial impact on sample size.

In Fig. 1, we explore in more detail the effects of misclas-

sification on sample size. The *solid lines* in Fig. 1 represent the sample size required to detect the specified 2-fold interaction in the absence of misclassification as a function of the true prevalence of the environmental factor for 0.5 (*Panels 1–3*) and 0.1 (*Panels 4–6*) prevalence of the genetic factor. The *dashed lines* in Fig. 1 illustrate the impact of misclassification of the environmental factor on sample size for selected values of sensitivity and specificity of exposure assessment.

For environmental and genetic factors with 0.5 true prevalence, reducing the environmental factor sensitivity from 1.0 to 0.8 and 0.6 (while holding specificity to 1.0) will increase the sample size from 720 to 1600 (2.2-fold) and to 3130 (4.4-fold) cases, respectively (Fig. 1, *Panel 1*). As the true prevalence of the environmental factor increases, the impact of reduced exposure assessment sensitivity will be stronger, as shown in Fig. 1, *Panel 1*. In contrast, reduced specificity has a stronger impact for rare than for common factors as shown in Fig. 1, *Panel 2*. For the specified parameters, reduced specificity tends to have a smaller impact on sample size than that of reduced sensitivity, except for very low prevalence of exposure. *Panel 3* shows the combined effect of reduced sensitivity and specificity of exposure assessment.

Fig. 1, *Panels 4–6* shows similar patterns as *Panels 1–3* for a genetic factor with 0.1 prevalence. For environmental factors with 0.5 true prevalence, reducing the environmental factor sensitivity from 1.0 to 0.8 and 0.6 (while holding specificity to 1.0) will increase the sample size from 1200 to 2700 (2.3-fold) and to 5390 (4.5-fold) cases, respectively. It should be noted that although the baseline sample size in the absence of misclassification is increased, the percent increase in sample size is very similar as it is in *Panels 1–3*. The reason is that in Fig. 1, we assumed that the genetic factor is perfectly measured and independent from the environmental factor. Therefore, the impact of misclassification on the environmental factor does not depend on the prevalence of the genetic factor.

Example of a 2-fold Additive Gene-Environment Interaction. Generally, when both the genetic and the environmental factors increase the risk of disease by themselves and in combination, as in the previous example, nondifferential misclassification tends to bias the additive interaction effect toward the null value. However, the direction of the bias to the additive interaction parameter due to nondifferential misclassification cannot be easily predicted. In this section, we provide an example of an additive gene-environment interaction where nondifferential misclassification of the environmental factor biases the additive interaction parameter away from the null value, even though the factors are
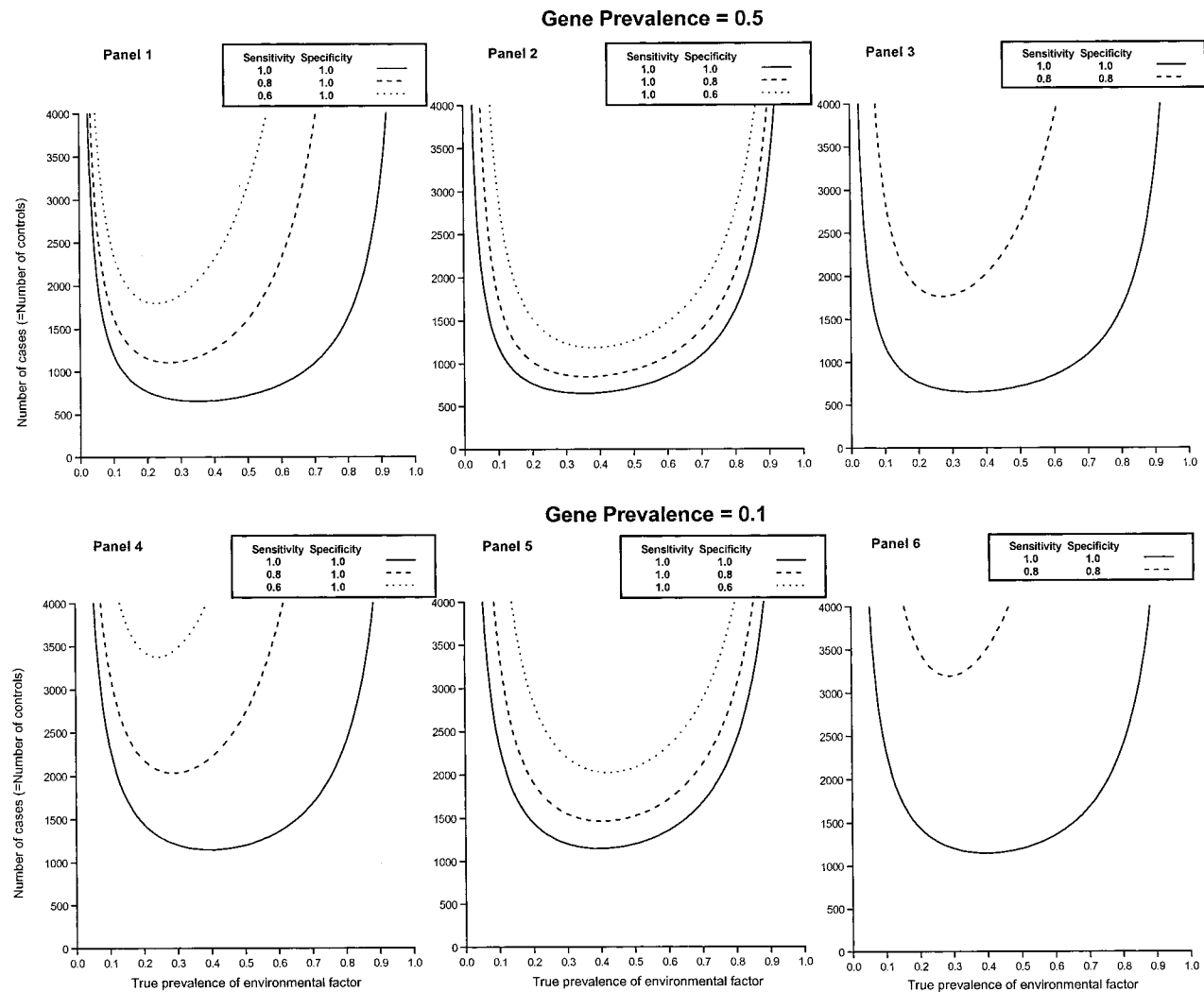
Gene Prevalence = 0.5

Gene Prevalence = 0.1

Fig. 1. Minimum number of cases (case:control ratio = 1) required to detect a 2-fold interaction ($OR_{10} = 2$, $OR_{01} = 2$, and $OR_{11} = 8$) with 80% power as a function of the true prevalence of the environmental factor ($P[E = 1]$), for 0.5 and 0.1 prevalence of the genetic factor, and for selected values of sensitivity and specificity of the environmental factor assessment.

Table 3 Minimum number of cases (case:control ratio = 1) required to detect a 2-fold additive gene-environment interaction ($OR_{10} = 0.5$, $OR_{01} = 2.0$, $OR_{11} = 2.0$) *for different levels of accuracy of the environmental and genetic factors[a]*

Environmental factor specificity	Genetic factor sensitivity	Prevalences		Φ^{*b}	Odds ratios			No. of cases
		Environmental factor	Genetic factor		OR_{10}	OR_{01}	OR_{11}	
1.0	1.0	0.30	0.50	2.00	0.5	2.00	2.00	2930
0.8	1.0	0.44	0.50	2.88	0.5	1.68	1.52	3486
1.0	0.95	0.30	0.48	1.87	0.51	2.05	2.05	3206
0.8	0.95	0.44	0.48	2.46	0.51	1.72	1.56	3663

[a] Sensitivity for environmental factor assessment = 1.0; specificity for genetic factor assessment = 1.0.
[b] Φ^*, observed interaction parameter.

binary and independent, and the misclassification probabilities for the environmental factor are independent of the genetic factor. In this example, the prevalence for the environmental factor is 0.3 and for the genetic factor is 0.5; the

odds ratio for the effect of the environmental factor alone (OR_{10}) is 0.5 and for the genetic factor alone (OR_{01}) is 2.0; and the joint odds ratio for both factors (OR_{11}) is 2.0. These values represent a 2-fold additive interaction ($\Phi = 2.0$).

Table 4 Minimum number of cases (case:control ratio = 1) required to detect a multiplicative interaction (Ψ)[a] between the COMT LL genotype and obesity in postmenopausal breast cancer risk for different levels of obesity accuracy

Method of obesity assessment	Obesity[b]			Ψ^*	No. of cases
	Sensitivity[c]	Specificity[c]	Prevalence		
Measured weight and height	1.0	1.0	0.15	2.00	1016
Self-reported weight and height	0.75	0.99	0.12	1.85	1548

[a] Assumptions and notation:
1. Ψ^*, observed interaction parameter.
2. Marginal odds ratio for obesity of 1.50 for measured weight and height and 1.43 for self-reported weight and height (26).
3. Prevalence of COMT LL of 25%; marginal odds ratio for COMT LL genotype of 2.00 (25).
4. COMT genotype is measured without error.
[b] BMI = weight/height^2 (kg/cm^2). Obesity: BMI $\geq$ 30.
[c] Estimates of sensitivity and specificity using measured weight and height as the gold standard (28).

Table 3 illustrates the impact of reducing the specificity of the environmental factor assessment from 1.0 to 0.8, both in the absence and presence of reduced sensitivity in the genetic factor assessment. The sensitivity for the environmental factor and the specificity for the genetic factor are both 1.0. In the absence of misclassification, the required sample size for 80% power is 2930 cases and 2930 controls. Reducing exposure specificity from 1.0 to 0.80 results in bias of the additive interaction parameter away from the null from 2.0 to 2.88 while increasing the required sample size to 3486 cases (1.19-fold increase). Paradoxically, when the sensitivity of the genetic factor is 0.95 rather than 1.0, the same amount of environmental factor error results in a smaller bias to the additive interaction parameter (from 2.0 to 2.46) while further increasing the required sample size to 3993 cases. Thus, reduced specificity in measuring a protective environmental factor can bias the additive interaction parameter away from the null value while increasing the sample size to reject the null hypothesis of no additive interaction.

COMT^2 Genotype, BMI, and Breast Cancer Risk. The COMT gene codes for an enzyme involved in the inactivation of estrogen catechols thought to be involved in breast carcinogenesis (23). A single-base polymorphism in the *COMT* gene has been associated with low-enzyme activity (24) and could result in decreased detoxification of catechol estrogens and subsequent increase in breast cancer risk (25). High BMI among postmenopausal women is associated with a moderate increase in breast cancer risk (26), which could be mediated by a higher production of estrogens among postmenopausal obese women (27). Thus, an investigator may want to evaluate if the odds ratio for obesity among postmenopausal women is higher for women with the *COMT LL* genotype (homozygous low activity) than for women with the *COMT HH* or *HL* genotypes (homozygous high activity or heterozygous).

Table 4 shows the minimum number of women needed to detect a 2-fold interaction (Ψ = 2.0) between obesity, defined as BMI $\geq$30 kg/m^2, and the *COMT LL* genotype using two alternative methods to estimate a women's BMI: with actual measurements of weight and height and self-reported weight and height. Assuming a prevalence of obesity of 0.15, a marginal odds ratio for obesity of 1.5, a prevalence of *COMT LL* genotype of 0.25, and a marginal odds ratio of 2.0 for *COMT*

LL genotype (25), one would need to study 1016 cases and 1016 controls to detect a 2-fold interaction (Ψ = 2.0). These marginal odds ratios and interaction parameter imply: OR_{10} = 1.10, OR_{01} =1.72, and OR_{11}= 3.78 (calculated as indicated in "Appendix 1").

If self-reported rather than actual measurements of weight and height are used to measure BMI, one would expect to classify correctly 75% women with truly high BMI and 99% with truly normal/low BMI (28). According to these misclassification probabilities, the observed interaction parameter, Ψ, will be 1.83 rather than 2.0, and the required sample size to detect the interaction will be increased from 1016 cases to 1548 cases and an equal number of controls. Although obtaining actual measurements of weight and height may increase the total cost of data collection, the savings from enrolling, collecting biological samples, and determining the genotype in 532 fewer cases and 532 fewer controls may off-set the increased cost of data collection. Moreover, using actual measurements of weight will provide unbiased estimates for the "true" interaction parameter and the obesity and *COMT LL* odds ratios.

Benzo(a)pyrene, GSTM1 Genotype, and Lung Cancer Risk Among Nonsmokers. Occupational exposure to benzo(a)pyrene has been associated with about a 2-fold increase in lung cancer risk among nonsmokers (29). Detoxification of benzo(*a*)pyrene by conjugation to glutathione is catalyzed by the GSTM1 enzyme (glutathione S-transferase M1). A homozygous deletion of the *GSTM1* gene is responsible for a lack of enzyme activity and has been associated with about a 1.5-fold increase in lung cancer risk (30). Thus, subjects exposed to benzo(*a*)pyrene who have the homozygous deletion in the *GSTM1* gene could be at a particularly high risk of lung cancer. Dewar *et al.* (31) have estimated a sensitivity of 0.6 and a specificity of 0.99 for the classification of exposure to benzo(*a*)pyrene based on a job-exposure matrix applied to job titles from a personal interview, as compared to exposure based on a more complex procedure involving the evaluation of a detailed job history by a trained team of chemists and industrial hygienists. Based on these estimates, a study to detect a 2-fold interaction (Ψ = 2.00, OR_{10} = 1.03, OR_{01} = 1.20, OR_{11}= 2.49) between the *GSTM1* null genotype and exposure to benzo(*a*)pyrene assessed by the evaluation of a detailed job history would need to include about 672 cases and 672 controls (Table 5). In contrast, using a job-exposure matrix to estimate benzo(*a*)pyrene exposure biases the interaction parameter to 1.76, and the required sample size is more than twice the previous estimate (1413 cases and 1413 controls).

---

[2] The abbreviations used are: COMT, catechol-O-metyltransferase; BMI, body mass index.

Table 5 Minimum number of cases (case : control ratio = 1) required to detect a multiplicative interaction[a] between lung cancer risk and occupational exposure to benzo(a)pyrene and the GSTM1 null genotype among nonsmokers for different levels of exposure accuracy

Method of benzo(a)pyrene exposure assessment	Benzo(a)pyrene			$\Psi*$	No. of cases
	Sensitivity[b]	Specificity[b]	Prevalence		
Detailed job history[c]	1.0	1.0	0.24	2.00	688
Job-exposure matrix[d]	0.60	0.99	0.15	1.76	1491

[a] Assumptions and notation:
1. $\Psi*$, observed interaction parameter.
2. Marginal odds ratio for benzo(a)pyrene exposure of 1.60 for detailed job history and 1.47 for job-exposure matrix (29).
3. Prevalence of GSTM1 null of 50%; marginal odds ratio for GSTM1 null genotype of 1.50 (30).
4. GSTM1 genotype is measured without error.
[b] Estimates of sensitivity and specificity using detailed job history as the gold standard (31).
[c] Substance exposure inferred from an in-depth interview of the subject with an evaluation of the subject's reported job history by a trained team of chemists and hygienists.
[d] Substance exposure inferred from a job-exposure matrix applied to the job titles obtained from an interview of the subject.

Discussion

Misclassification of environmental or genetic risk factors can greatly increase the sample size required to evaluate gene-environment interactions in case-control studies. As illustrated in our examples, when the interaction effect is moderate to small, even relatively small biases to the interaction parameter can lead to large increases in sample size. This is because sample size requirements tend to increase nonlinearly as effects become closer to the null value. The effects of misclassification are highly dependent on both the true risk model and the distribution of the misclassified risk factors in the population; therefore, the potential effects of misclassification on the interaction parameter and the required sample size should be evaluated in each particular situation. This paper provides a procedure to determine the observed interaction parameter and required sample size based on assumptions about the accuracy of exposure assessment. This procedure can be used for any pattern of multiplicative or additive gene-environment interaction such as those described by Ottman *et al.* (22) and Khoury *et al.* (21). Moreover, this procedure is not unique to studies involving genetic factors, and it can be used for any two binary and independent factors, measured with errors that are independent of each other. More complex procedures are needed for polytomous categorical or continuous factors, for non-independent factors, or for factors measured with correlated errors.

Both differential and nondifferential misclassification of the environmental factor biases a multiplicative interaction effect toward the null value provided that the environmental and genetic factors are binary and independent, misclassification is independent of the genetic factor, and the sum of sensitivity and specificity is ≥1 (*i.e.,* the classification instrument is better than random; Ref. 11). However, bias to the additive interaction parameter cannot be easily predicted, even under this set of conditions. In fact, we provide an example of an additive interaction between a genetic susceptibility factor and a protective environmental factor, where reduced specificity in the assessment of the environmental factor results in an overestimation of the additive interaction parameter and an increase in sample size. Although this and all other examples in this paper assume nondifferential misclassification with respect to the disease status, our procedure can also be used for differential misclassification.

The observations in this paper point out the trade-off between using more accurate and usually more expensive measures of exposure assessment in a smaller number of subjects or using less-accurate but usually cheaper measures in a larger number of subjects. When making these choices, it should be borne in mind that increasing sample size increases the study power to detect the attenuated interaction; however, the interaction effect is still biased. In this case, adjustments based on estimates of sensitivity and specificity are required to obtain an unbiased estimate of the true interaction effect. It should be noted that if the conditions used in our paper are *not* satisfied (*i.e.,* binary genetic and environmental factors, independent of each other in the population and independence of misclassification probabilities for both factors), there may be unpredictable effects of misclassification on the direction of the bias to the multiplicative interaction (11). Moreover, as indicated above, the direction of the bias to the additive interaction cannot be generally predicted, even under the conditions used in our paper.

In conclusion, efforts to improve the accuracy of exposure assessment for both the environmental and genetic factors can greatly reduce sample size requirements to study interactions and are critical for accurate assessment of gene-environment interactions in case-control studies. Our examples also illustrate the importance of routine assessment of accuracy in genotype assays through quality control procedures because of the large impact of small degrees of error.

Appendix 1
Calculation of Marginal Odds Ratios

For given values of the effects for the environmental and genetic factor alone (OR_{10} and OR_{01}), interaction effect (ψ or ϕ), and prevalence of the environmental factor and genetic factors ($P(E = 1)$ and $P(G = 1)$), the environmental and genetic marginal effects (OR_E and OR_G), can be calculated by using in the following set of equations.

For multiplicative interactions:

$$OR_E = \frac{(1 - P(G = 1))*OR_{10} + P(G = 1)*OR_{10}*OR_{01}*\Psi}{(1 - P(G = 1)) + P(G = 1)*OR_{01}}$$

$$OR_G = \frac{(1 - P(E = 1))*OR_{01} + P(E = 1)*OR_{10}*OR_{01}*\Psi}{(1 - P(E = 1)) + P(E = 1)*OR_{10}}$$

For additive interactions:

$$OR_E = \frac{(1 - P(G = 1))*OR_{10} + P(G = 1)*[(OR_{10} + OR_{01} - 2)*\Psi + 1]}{(1 - P(G = 1)) + P(G = 1)*OR_{01}}$$

$$OR_G = \frac{(1 - P(E = 1))*OR_{01} + P(E = 1)*[(OR_{10} + OR_{01} - 2)*\Psi + 1]}{(1 - P(E = 1)) + P(E = 1)*OR_{10}}$$

For given values of the environmental and genetic marginal effects (OR_E and OR_G), interaction effect (Ψ or Φ), and prevalence of the environmental factor and genetic factors ($P(E = 1)$ and $P(G = 1)$), the effects of the environmental and genetic factors alone (OR_{10} and OR_{01}) can be calculated by solving for OR_{10} and OR_{01} in the above set of equations.

All calculations in this Appendix can be performed easily using a spreadsheet (EXPECT) that can be obtained by e-mail from connorj@mail.nih.gov.

Appendix 2: Calculation of observed parameters in the presence of misclassification

The observed parameters in the presence of misclassification, $P(G^* = 1)$, $P(E^* = 1)$, OR^*_{10}, OR^*_{01}, and OR^*_{11}, for given values for sensitivity and specificity can be calculated using a spreadsheet (EXPECT) that can be obtained by e-mail from connorj@mail.nih.gov.

Below are the formulae used in calculations performed by EXPECT.

Expected cell counts in the absence of misclassification from a case-control study of a gene-environment interaction

	$G = 1$			$G = 0$	
	$E = 1$	$E = 0$		$E = 1$	$E = 0$
$D = 1$	a_1	b_1	$D = 1$	a_0	b_0
$D = 0$	c_1	d_1	$D = 0$	c_0	d_0

Given $P(G = 1)$, $P(E = 1)$, OR_{10}, OR_{01}, and OR_{11}, the expected cell counts in Table 2 are:

$a_1 = \lambda \times P(E = 1) \times P(G = 1) \times OR_{11}$
$b_1 = \lambda \times (1 - P(E = 1)) \times P(G = 1) \times OR_{01}$
$c_1 = P(E = 1) \times P(G = 1)$
$d_1 = (1 - P(E = 1)) \times P(G = 1)$
$a_0 = \lambda \times P(E = 1) \times (1 - P(G = 1)) \times OR_{10}$
$b_0 = \lambda \times (1 - P(E = 1)) \times (1 - P(G = 1))$
$c_0 = P(E = 1) \times (1 - P(G = 1))$
$d_0 = (1 - P(E = 1)) \times (1 - P(G = 1))$

where $\lambda = \dfrac{1}{(1 - P(E=1)) \times (1 - P(G=1)) + P(E=1) \times (1 - P(G=1)) \times OR_{10} + (1 - P(E=1)) \times P(G=1) \times OR_{01} + P(E=1) \times P(G=1) \times OR_{11}}$

Let $se_{0E}\ sp_{0E}$ and $se_{1E}\ sp_{1E}$ be the sensitivity and specificity of the environmental factor among controls and cases, respectively, and $se_{0G}\ sp_{0G}$ and $se_{0G}\ sp_{0G}$ be the sensitivity and specificity of the genetic factor among controls and cases, respectively. The expected cell counts among the controls in the presence of misclassification (denoted by an asterisk *) are calculated as:

$$
\begin{bmatrix}
se_{0E} \times se_{0G} & (1 - sp_{0E}) \times se_{0G} & se_{0E} \times (1 - sp_{0G}) & (1 - sp_{0E}) \times (1 - sp_{0G}) \\
(1 - se_{0E}) \times sg_{0G} & sp_{0E} \times se_{0G} & (1 - se_{0E}) \times (1 - sp_{0G}) & sp_{0E} \times (1 - sp_{0G}) \\
se_E \times (1 - se_G) & (1 - sp_{0E}) \times (1 - se_{0G}) & se_{0E} \times sp_{0G} & (1 - sp_{0E}) \times sp_{0G} \\
(1 - se_{0E}) \times (1 - se_{0G}) & sp_{0E} \times (1 - se_{0G}) & (1 - se_{0E}) \times sp_{0G} & sp_{0E} \times sp_{0G}
\end{bmatrix}
\times
\begin{bmatrix} c_1 \\ d_1 \\ c_0 \\ d_0 \end{bmatrix}
=
\begin{bmatrix} c_1^* \\ d_1^* \\ c_0^* \\ d_0^* \end{bmatrix}
$$

The expected cell counts among cases are:

$$
\begin{bmatrix}
se_{1E} \times se_{1G} & (1 - sp_{1E}) \times se_{1G} & se_{1E} \times (1 - sp_{1G}) & (1 - sp_{1E}) \times (1 - sp_{1G}) \\
(1 - se_{1E}) \times se_{1G} & sp_{1E} \times se_{1G} & (1 - se_{1E}) \times (1 - sp_{1G}) & sp_{1E} \times (1 - sp_{1G}) \\
se_{1E} \times (1 - se_{1G}) & (1 - sp_{1E}) \times (1 - se_{1G}) & se_{1E} \times sp_{1G} & (1 - sp_{1E}) \times sp_{1G} \\
(1 - se_{1E}) \times (1 - se_{1G}) & sp_{1E} \times (1 - se_{1G}) & (1 - se_{1E}) \times sp_{1G} & sp_{1E} \times sp_{1G}
\end{bmatrix}
\times
\begin{bmatrix} a_1 \\ b_1 \\ a_0 \\ b_0 \end{bmatrix}
=
\begin{bmatrix} a_1^* \\ b_1^* \\ a_0^* \\ b_0^* \end{bmatrix}
$$

The observed parameters in the presence of misclassification, $P(G^* = 1)$, $P(E^* = 1)$, OR^*_{10}, OR^*_{01}, and OR^*_{11}, are then calculated from the expected cell counts. Note that for nondifferential misclassification of the environmental and genetic factors $se_{0E} = se_{1E}$, $sp_{0E} = sp_{1E}$, and $se_{0G} = se_{1G}$, $sp_{0G} = sp_{1G}$, respectively.

References

1. Bross, I. Misclassification in 2x2 tables. Biometrics, 478–489, 1954.

2. Diamond, E. L., and Lilienfield, A. M. Effects of errors in classification and diagnosis in various types of epidemiologic studies. Am. J. Public Health, *52:* 1137–1144, 1962.

3. Copeland, K. T., Checkoway, H., McMichael, A. J., and Holbrook, R. H. Bias due to misclassification in the estimation of relative risk. Am. J. Epidemiol., *105:* 488–495, 1977.

4. Greenland, S. The effect of misclassification in the presence of covariates. Am. J. Epidemiol., 1980; 564–569.

5. Flegal, K. M., Brownie, C., and Haas, J. D. The effects of exposure misclassification on estimates of relative risk. Am. J. Epidemiol., *123:* 736–750, 1986.

6. Armstrong, B. K., White, E., and Saracci, R. Principles of Exposure Measurement in Epidemiology. Oxford: Oxford University Press, 1992.

7. Quade, D., Lachenbruch, P. A., Whaley, F. S., McClish, D. K., and Haley, R. W. Effects of misclassification on statistical inferences in epidemiology. Am. J. Epidemiol., *111:* 503–515, 1980.

8. Rothman, N., Stewart, W. F., Caporaso, N. E., and Hayes, R. B. Misclassification of genetic susceptibility biomarkers: implications for case-control studies and cross-population comparisons. Cancer Epidemiol. Biomark. Prev., *2:* 299–303, 1993.

9. Dosemeci, M., Wacholder, S., and Lubin, J. H. Does nondifferential misclassification of exposure always bias a true effect toward the null value? Am. J. Epidemiol., *132:* 746–748, 1990.

10. Rothman, N., Garcia-Closas, M., Stewart, W. F., and Lubin, J. H. The impact of misclassification in studies of gene-environment interactions. *In:* Vineis P., Malats N., Lang M., *et al.*, eds. Metabolic polymorphisms and susceptibility to cancer, 89–98. Lyon: IARC, 1999.

11. Garcia-Closas, M., Thompson, W. D., and Robins, J. M. Differential misclassification and the assessment of gene-environment interactions in case-control studies. Am. J. Epidemiol., *147:* 426–433, 1998.

12. Smith, P. G., and Day, N. E. The design of case-control studies: the influence of confounding and interaction effects. Int. J. Epidemiol., *13:* 356–365, 1994.

13. Hwang, S-J., Beaty, T., Liang, K-Y., Coresh, J., and Khoury, M. J. Minimum sample size estimation to detect gene-environment interaction in case-control designs. Am. J. Epidemiol., *140:* 1029–1037, 1994.

14. Flanders, W. D., and Khoury, M. J. Analysis of case-parental control studies: method for the study of associations between disease and genetic markers. Am. J. Epidemiol., *144:* 696–703, 1996.

15. Goldstein, A., Falk, R. T., Korczak, J. F., and Lubin, J. H. Detecting gene-environment interactions using a case-control design. Genetic Epidemiol., *14:* 1085–1089, 1997.

16. Garcia-Closas, M., and Lubin, J. H. Power and sample size calculations in case-control studies of gene-environment interactions: comments on different approaches. Am. J. Epidemiol., 689–692, 1999.

17. Rothman, K., Greenland, S., and Walker, A. M. Concepts of interaction. Am. J. Epidemiol., *112:* 467–470, 1998.

18. Kleinbaum, D. G., Kupper L. L., and Morgenstern, H. Epidemiologic Research: Principles and Quantitative Methods. New York: Van Nostrand Reinhold, 1982.

19. Lubin, J. H., and Gail, M. On power and sample size for studying features of the relative odds of disease. Am. J. Epidemiol., *131:* 552–566, 1990.

20. Lubin, J. H., Boice, J. D., Jr., and Samet, J. M. Errors in exposure assessment, statistical power and the interpretation of residential radon studies. Radiat. Res., *144:* 329–341, 1995.

21. Khoury, M. J., Beaty, T. H., and Cohen, B. H. Fundamentals of Genetic Epidemiology. Oxford: Oxford University Press, 1993.

22. Ottman, R., Pike, M. C., and King, M-C. Familial breast cancer in a population-based series. Am. J. Epidemiol., *123:* 15–21, 1986.

23. Yager, J. D., and Liehr, J. G. Molecular mechanisms of estrogen carcinogenesis. Annu. Rev. Pharmacol. Toxicol., *36:* 203–232, 1996.

24. Scanlon, P. D., Raymond, F. A., and Weinshilboum, R. M. Cathecol-O-methyltransferase: thermolabile enzyme in erythrocytes of subjects homozygous for allele for low activity. Science (Washington DC), *203:* 60–65, 1979.

25. Lavigne, J. A., Helzlsouer, K., and Huang, H-Y., *et al.* An association between the allele coding for a low activity variant of catechol- *O*-methyltransferase and the risk of breast cancer. Cancer Res., *57:* 5493–5497, 1997.

26. Hunter, D. J., and Willet, W. Diet, body size, and breast cancer. Epidemiol. Rev., *15:* 110–132, 1993.

27. Feigelson, H. S., and Henderson, B. E. Estrogens and breast cancer. Carcinogenesis, *17:* 2279–2284, 1996.

28. Nieto-Garcia, F. J., Bush, T. L., and Kelyl, P. M. Body mass definitions of obesity: sensitivity and specificity using self-reported weight and height. Epidemiology, *1:* 146–152, 1990.

29. Nadon, L., Siemiatycki, J., Dewar, R., Krewski, D., and Gerin, M. Cancer risk due to occupational exposure to polycyclic aromatic hydrocarbons. Am. J. Ind. Med., *3:* 303–324, 1995.

30. McWilliams, J. E., Sanderson, B. J. S., Harris, E. L., Richert-Boe, K. E., and Henner, W. D. Glutathione S-transferase M1 (GSTM1) deficiency and lung cancer risk. Cancer Epidemiol. Biomark. Prev., *5:* 589–594, 1995.

31. Dewar, R., Siemiatycki, J., and Gerin, M. Loss of statistical power associated with the use of a job-exposure matrix in occupational case-control studies. Appl.Occup. Environ. Hyg., *6:* 508–515, 1997